Taiwanese Across Taiwan Corpus - TAT-Vol1 and TAT-Vol2

Yuan-Fu Liao
National Taipei University of Technology
yfliao@mail.ntut.edu.tw

1. Introduction

TAT (Taiwanese across Taiwan) corpus is a large-scale multi-channel read-speech corpus recorded across Taiwan using 6 different microphones in quite office-like environment. It contains 300 hours x 6 channels speech produced by 600 speakers. The first two volumes of TAT corpus, TAT-Vol1 and TAT-Vol2, in total 200 speakers, about 100 hours, have been well-transcribed and therefore publicly released.

2. Corpus Design

The following strategy was chosen to alleviate labelling efforts:

- Recruit native Taiwanese speakers across Taiwan to cover regional variations
- Adapt the official Taiwanese Romanization system proposed by Taiwan Ministry of Education (MOE) as the writing standard
- Record reading speech with prepared prompt sheets instead of transcribing spontaneous speech by linguists

3. Recording Protocol

- Text Materials and Prompt Sheets: The adopt Taiwanese native text articles mainly came from articles published by two sources:
  - Li Kang Khioh Taiwanese Cultural and Educational Foundation (李江台語文教基金會) [14]: 50 authors, about 6,000 words per author, and a daily conversation textbook with 14 lessons
  - MOE: 250 articles, about 600 words per article

Example:

1
運動顧健康
ūn-tōng kòo kiān-khong

2
Tsiânn久無見面，你看--起來有khah瘦，
tsiânn kú bô kìnn-bīn, lí khuànn-khí-lâi ū khah sán,

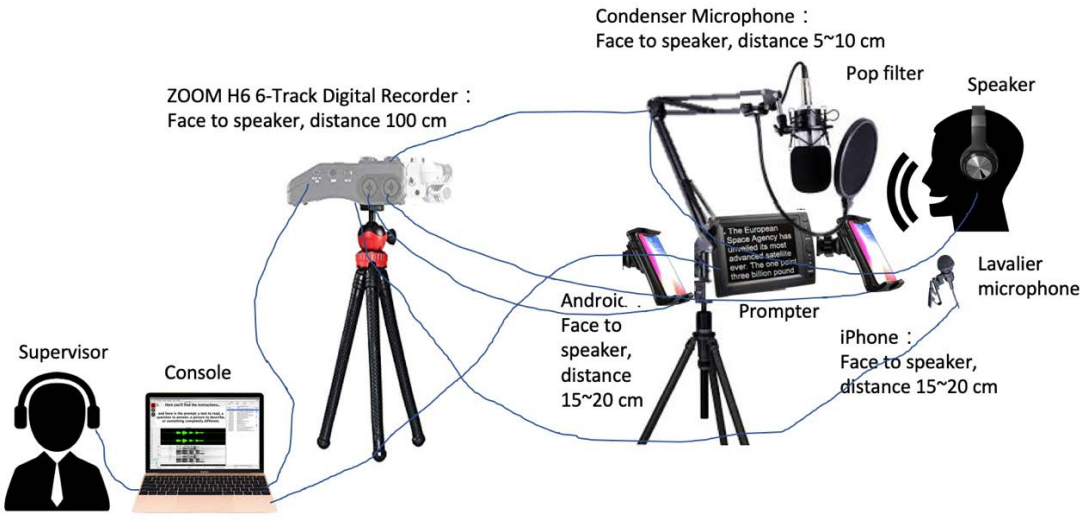
3
koh比進前ke tsiok有元氣！
koh pí tsìn-tsîng ke tsiok ū guân-khì!

4
瘦是m̄-tsiânn瘦--lah,
sán sī m̄-tsiânn sán-lah,

5
幾若個月tsiah瘦兩三公斤niâ。
kuí-nā/kuí-lō/kuí-ā kò guèh/gèh tsiah sán nn̄g sann kong-kin/kong-kun niâ.

6
Hóo！兩三公斤就真gâu--ah,
hóo! nn̄g sann kong-kin/kong-kun tō/tiō/tiòh/tòh tsin gâu-ah,

- Configurations: Six different microphones were adopted at the same time for data collection in a quiet office-like environment
  - Recording Software: SpeechRecorder
  - Microphones
    - Digital audio interface: ZOOM H6
    - Close-talk: Audio-Technica AT2020
    - Lavalier: Superlux WO518+PS418D
    - Distant: ZOOM XYH-6 stereo microphone
    - iPhone App: Microphone Live
    - Android Phone App: Microphone



- Audio file format: WAVE, 16 kHz sampling rate, 16 bits PCM
- Transcription file format: JSON

0011-1.1.json 823 Bytes

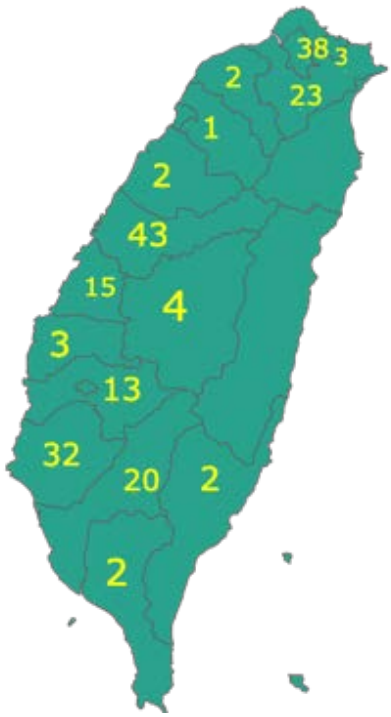

```
{
    "音檔長度": "8.59",
    "漢羅台文": "我厝內的電話是空二三三六六九空五四",
    "台羅": "guá tshù-lāi ê tiān-uē sī khòng lī sam sam liok liok kiú khòng ngóo sù",
    "台羅數字調": "gua2 tshu3-lai7 e5 tian7-ue7 si7 khong3 li7 sam1 sam1 liok8 liok8 kiu2 khong3 ngoo2 su3",
    "白話字": "góa chhù-lāi ê tiān-ōe sī khòng lī sam sam liok liok kiú khòng ngó sù",
    "字數": "17",
    "提示卡編號": "0011",
    "句編號": "1.1",
    "發音人": "IUF003",
    "性別": "女",
    "年齡": "61",
    "教育程度": "國中",
    "出生地": "台南市南區",
    "現居地": "台南市南區",
    "腔調": "台南混合腔",
    "錄音環境": "安靜隔音室內",
    "提示卡切換速度": "快",
    "總錄音時間(分)": "90"
}
```


4. Statistics

- Content of TAT-Vol1~2 including number of speakers, sentences and characters and total speech duration in hours.

TAT-Vol1				
	Speakers	**Sentences**	**Characters**	**Hours**
Train	80	23,104	271,772	41.76
Evaluation	10	2,943	34,426	5.02
Test	10	2,786	33,394	5.16
TAT-Vol2				
	Speakers	**Sentences**	**Characters**	**Hours**
Train	80	23,216	272,671	42.39
Evaluation	10	2,951	35,951	4.76
Test	10	2,811	31,985	5.27
Total				
	Speakers	**Sentences**	**Characters**	**Hours**
Total	200	57,811	680,199	104.36

- Distribution of Speakers' Residences



- Distribution of Speakers' Ages